A Survey on Biomedical Text Summarization with Pre-trained Language Models

Qianqian Xie, Zheheng Luo, Jimin Huang, Hua Wang, Sophia Ananiadou

Abstract—Biomedical text summarization has long been a fundamental task in biomedical natural language processing (BioNLP), aiming at generating concise summaries that distill key information from one or multiple biomedical documents. In recent years, pre-trained language models (PLMs) have been the de facto standard of various natural language processing tasks in the general domain. Most recently, PLMs have been further investigated in the biomedical domain and brought new insights into the biomedical text summarization task. In this paper, we systematically summarize recent advances that explore PLMs for biomedical text summarization. We categorize PLMs-based approaches according to types of input documents, numbers of input documents, and types of output summary. We review available datasets, recent approaches and evaluation methods of the task. We finally discuss existing challenges and promising future directions. We hope the survey will be useful for researchers in the research community to quickly track recent progress and provide guidelines for future research.

Index Terms—Biomedical texts, text summarization, pre-trained language models.

1 Introduction

With the rapidly increasing of unstructured clinical information, such as biomedical literature [1] and clinical notes [2], it is quite challenging for researchers and clinicians to access the required information quickly. To meet the challenge, the text summarization technique [3] has been explored in the biomedical domain to help users seek information more efficiently. Biomedical text summarization [4] aims to shorten one or multiple biomedical documents into a condensed summary that keeps the most important semantic information. It saves much time and effort for users since they can grasp the main idea of long biomedical documents by only reading the summary. It can be applied in various real applications including but not limited to aiding evidence-based medicine [5], clinical information management [6], and clinical decision support [7].

Throughout these years, biomedical text summarization has been facilitated by methods widely used in the text summarization of the general domain, including graph-based ranking methods [8], traditional machine learning methods [9] and deep learning methods [10]. Compared with unstructured texts in the general domain, characteristics of unstructured biomedical texts such as complex language structures [11] and laden with jargon, bring more challenges for automated text summarization methods. Therefore, it requires automated text summarization methods to capture the biomedical domain knowledge to understand biomedical texts. In recent years, pre-trained language models (PLMs) [12] that are pre-trained on large-scale of unlabeled texts self-supervisedly, have been the paradigm of various natural language processing tasks. Self-supervised pre-training makes PLMs memorize common sense and lexical knowledge inherited in the training texts [13], which can be then transferred to improve NLP tasks via fine-tuning. Without manually annotated data, PLMs can greatly boost the performance of various NLP tasks via knowledge transfer. Thus, they are quite friendly for the scenario where large-scale unlabeled data is available but lacks human-labeled data, such as the biomedical domain. Motivated by the great success of PLMs, many efforts [14], [15] have been proposed to explore PLMs for the biomedical text summarization task, and greatly improve the performance of the task recently. These methods consider general-domain language models such as BERT [12] or domain-specific language models such as BioBERT [16], as the backbone model for encoding input texts. And then they are fine-tuned with the specific loss and dataset of the biomedical text summarization task, which allows the semantic knowledge captured in PLMs to be transferred to the summarization task.

Although there were surveys on biomedical text summarization including the earliest ones [17], [4], [18], and the most recent one [19], they described existing methods that utilized traditional machine learning and deep learning techniques rather than methods based on PLMs. However, with the quick development of PLMs for biomedical text summarization, it requires a comprehensive survey of recent publications since the era of PLMs to help track the recent progress. To fill the gap, this paper surveys recent work that utilizes PLMs for the biomedical text summarization task. We systematically review recent approaches, benchmark datasets and evaluation methods of the task. We categorize and discuss existing methods according to types of input documents: including biomedical literature, clinical texts, medical dialogues, and medical questions, numbers of input documents: single document or multiple documents, and types of output summary: extractive or abstractive. We hope this paper can be a timely survey for researchers in the research community to quickly track recent progress. The main contributions of this survey are:

- We propose a comprehensive review of biomedical

• *Qianqian Xie is with the Department of Computer Science, University of Manchester, Manchester, United Kingdom.*
E-mail: qianqian.xie@manchester.ac.uk

text summarization with pre-trained language models. To the best of our knowledge, this is the first review that surveys recent PLMs-based methods.
- We categorize and discuss recent approaches, benchmark datasets and evaluation methods thoroughly.
- We discuss challenges of existing approaches and outlook promising future directions.

Compared with existing surveys [17] was the earliest survey that summarized traditional natural language processing and machine learning methods for medical document summarization. [4] reviewed text summarization methods for biomedical literature and electronic health records (EHRs), between January 2000 and October 2013. [18] examined automated summarization methods for electronic health records. Most methods summarized in these surveys are traditional machine learning methods based on feature engineering. With the prosperity of deep learning since 2014, deep neural networks became the mainstream method for biomedical text summarization. Recently, [19] investigated text summarization approaches for both biomedical literature and EHRs between January 2013 to April 2021. Although it further reviewed deep learning-based methods, methods with PLMs were not investigated in the survey. [20] surveyed recent progress of pre-trained language models for the biomedical domain, in which PLMs-based methods for various biomedical natural language processing tasks are introduced, including the biomedical text summarization task. However, it only introduced recent approaches briefly and didn't provide a comprehensive overview of the task, such as benchmark datasets, evaluation methods, limitations et al.

Paper collection We collect representative works since 2018 that are published in conferences and journals of computer science and biomedical science such as ACL, EMNLP, COLING, NAACL, AAAI, Bioinformatics, BioNLP, JAMIA, AMIA et al. We use google scholar as the search engine, and search with keywords including "biomedical summarization", "medical summarization", "clinical summarization", "medical dialogue summarization" et al.

Organization of the paper We will first introduce the background of biomedical text summarization and pre-trained language models in the Section 2. Then Section 3 will describe benchmark datasets. Representative PLMs based methods will be categorized and discussed in Section 4. We introduce evaluation methods in Section 5. We next discuss limitations and future directions in Section 6. Finally, we make a conclusion in Section 7. Figure 1 shows the proposed overview of biomedical text summarization with pre-trained language models.

2 Background

In this section, we first review biomedical text summarization and pre-trained language models, which are two essential concepts used in this survey. The overview of the background section is shown in Figure 2.

2.1 Biomedical Text Summarization

Biomedical text summarization aims to shorten single or multiple biomedical documents into a condensed summary that preserves the most important information from the original text. In general, automated summarization approaches are divided into extractive summarization methods [21] and abstractive summarization methods [22] according to the output of summaries. Extractive methods select key sentences from original documents and concatenate them into a summary, while abstractive methods generate new sentences as the summary based on the original documents.

Automatic biomedical summarization methods are largely facilitated and insipired by automatic methods in general domain. The earliest methods are traditional machine learning methods such as Naive-Bayes classifier [23], and graph based ranking methods such as TextRank [8]. With the prosperity of deep learning since 2014, neural network methods have been the mainstream method for both extractive and abstractive summarization of biomedical texts. Neural extractive methods [21] formulate the extractive task as the binary classification problem that predicts labels (1 or 0) of sentences in original documents to select sentences. Neural abstractive methods [24] model the abstractive task as the text generation problem that generates new sentences based on the sequence-to-sequence framework. Compared with extractive summarization, abstractive summarization is more challenging. It is difficult for automated abstractive methods to generate factually consistent summaries, since it involves generating informative sentences from the large vocabulary, lexical and syntactic adjustment, and paraphrasing. Formally, let's assume C as a biomedical corpus with D documents, $d \in C$ is a document consisting of m sentences: $d = \{s_1, \cdots, s_m\}$. We also assume the gold summary of the document d as t_d. For most scientific paper datasets, abstracts of papers are deemed as their gold summaries.

Extractive summarization For document d, extractive summarization methods aim to select a subset of o sentences from d, $o \ll m$. Neural extractive methods can be classified into unsupervised methods and supervised methods. The unsupervised methods model the extractive task into the sentence ranking problem. They generate sentence representations based on word embeddings, and use the unsupervised ranking method to select important sentences based on their representations. For supervised methods, unsupervised sentence selection methods such as the greedy search algorithm [21] are firstly used to extract binary labels for sentences of each document, that are used to train the neural extractive models. The oracle summary includes sentences with label 1, that has the maximum semantic similarity with the gold summary of each document. While remaining sentences with label 0 will not be included in the oracle summary. Most of neural extractive methods consists of the neural network based encoder and classifier. Neural network based encoder is used to capture the contextual information of input documents and generate vector representations of sentences. The classification layer is to predict labels of sentences with their vector representations. The objective is to maximize the log-likelihood of the observed labels of sentences:

$$\log p(y|C;\theta) = \sum_{d \in C} \sum_{i=1}^{m} \log p(y_i^d | d; \theta) \quad (1)$$

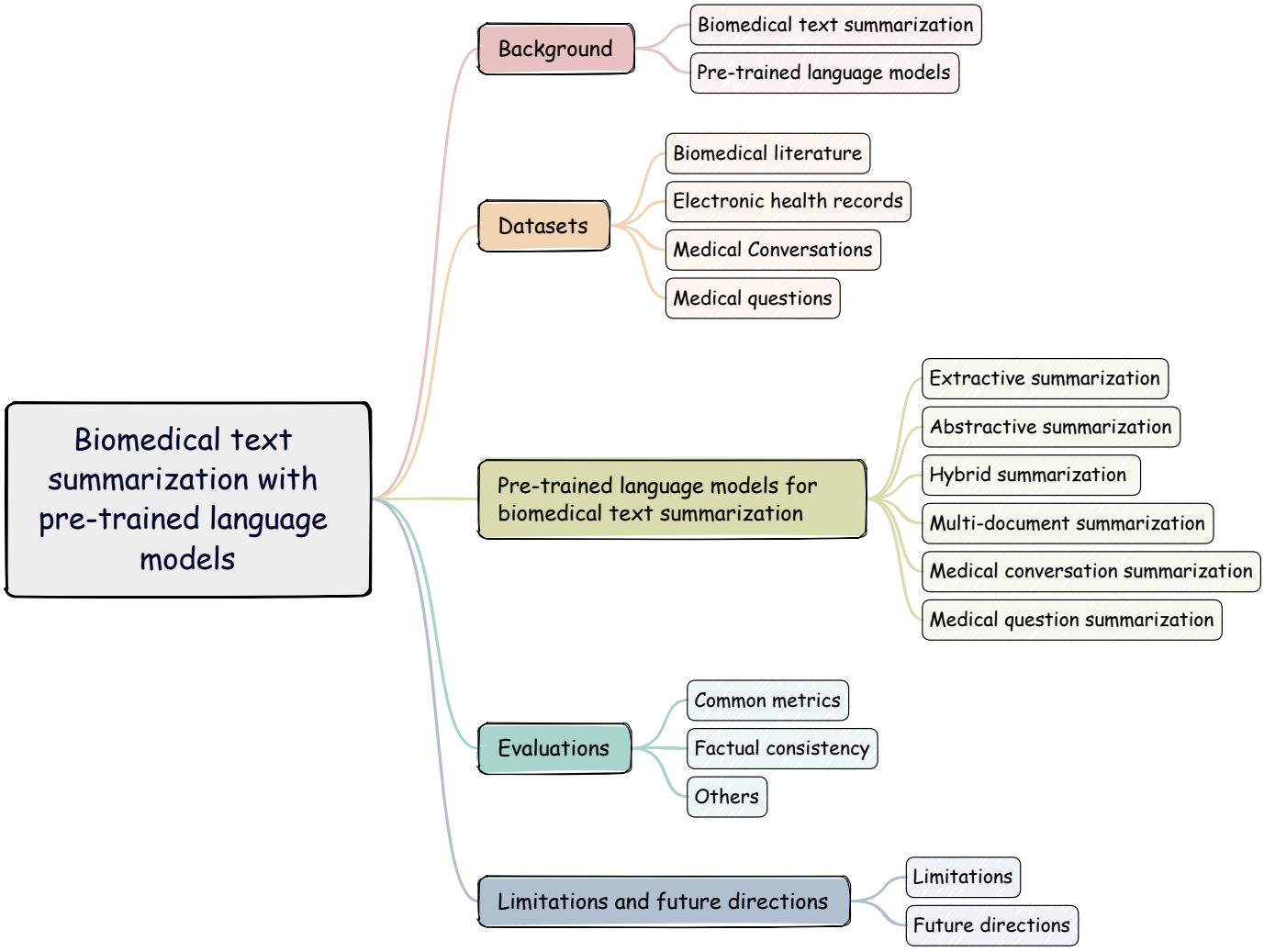


Fig. 1. Overview of biomedical text summarization with pre-trained language models.

where y_i^d is the ground truth label of sentence s_i in document d, θ is the parameter set of the model.

Abstractive summarization Neural abstractive methods build the abstractive task as the sequence-to-sequence learning problem. Most of them utilize the encoder-decoder framework [25], that consists of the neural network based encoder and decoder. Similar to extractive methods, the encoder is used to yield vector representations of input documents. The decoder is to generate the target summary sequentially with representations from the encoder. The model is optimized via the objective to maximize the log-likelihood of target words in the gold summary.

$$\log p(t|C;\theta) = \sum_{d \in C} \sum_{i=1}^{n} \log p(t_i^d | d; \theta) \tag{2}$$

where t_i^d is the i-th word in the gold summary t^d of the document d, $n \ll m$.

Most recently, PLMs have become the new paradigm of biomedical summarization. PLMs based methods have the similar framework with neural methods on extractive and abstractive task, while PLMs are more powerful than neural networks on encoding biomedical texts. We will next introduce the pre-trained language models.

2.2 Pre-trained Language Models

Language model pre-training [26], [27] has long been an active research area with the aim of learning low-dimensional vector representations from natural language, which are applicable and generalizable for downstream tasks. The earliest unidirectional neural language models such as word2vec [27], glove [28], learn meaningful word embeddings via estimating the probability of next word with the sequence of history words. The bidirectional language models such as ELMo [29] are then proposed to further consider the bidirectional context of words. Bidirectional Encoder Representations from Transformers (BERT) [12] is the breakthrough work that advances the state-of-art of various NLP tasks. It proposes to first pre-train the deep models based on basic neural network structure such as transformer [30] on large scale of unlabeled data with self-supervised learning task. And then pre-trained parameters of deep models and task-specific parameters are fine-tuned on labeled data with down stream tasks. Recently, BERT and its variants have greatly facilitated performance of natural language processing tasks. The two steps: pre-training and fine-tuning has become the paradigm of NLP tasks. Model architecture, pre-training, training corpora and fine-tuning

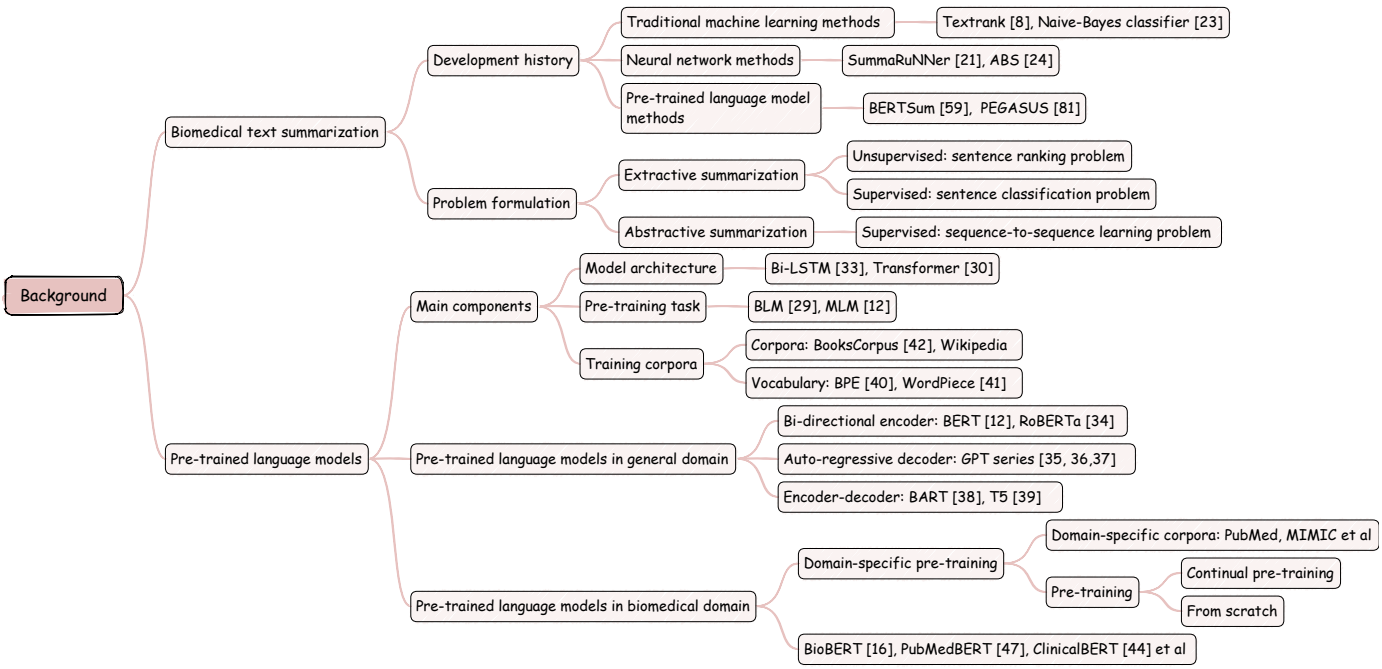


Fig. 2. Overview of background.

are core technical components of PLMs. One can check the review [31] for more details of PLMs.

Model architecture The early language models such as ELMo and its predecessors [29], [32], generally utilize Bi-LSTM [33] as the basis network structure, to capture bi-directional contextual information of texts. However, Bi-LSTM has the limitation of parallelization and sequential computation with the growing of sequence length. One breakthrough work is Transformer [30], that proposes the self-attention based neural network model architecture, that is able to parallel computation and model long-range dependencies of sequences efficiently. The Transformer follows the encoder-decoder architecture with stacked multi-head self-attention and point-wise fully connected feed-forward network. After that, nearly all pre-trained language models utilize the Trasformer architecture. According to different model architecture, existing PLMs can be categorized into three types: masked language models, auto-regressive lanaguage model, and encoder-decoder language models. The masked language models, such as BERT and its variants such as Roberta [34] use Transformer as the bi-directional encoder. The auto-regressive language models such as GPT series [35], [36], [37], only pre-train auto-regressive decoder based on Transformer architecture. While encoder-decoder language models such BART [38] pre-train the full encoder-decoder Transformer architecture. To learn better representations, they usually have deep network architecture. For example, the base model of BERT has 12 Transformer layers with hidden size 768 and 12 self-attention heads.

Pre-training The pre-training task on large scale of unlabeled data is the key for language models to learn useful representations and parameters, which can be fine-tuned to down stream tasks. The pre-training task of most previous language models follows the unidirectional language model [26]. It aims to maximize the log-likelihood of words conditionally on history words:

$$\mathcal{L}_{lm} = -\sum_{t=1}^{T} \log p(x_t|x_1, x_2, \cdots, x_{t-1}) \tag{3}$$

where $X = \{x_1, \cdots, x_T\}$ is a given text sequence. The bidirectional language model is further proposed to capture contextual information of text from both directions. It combines both the left-to-right language model and righ-to-left language model:

$$\begin{aligned}\mathcal{L}_{blm} = -\sum_{t=1}^{T} (&\log p(x_t|x_1, x_2, \cdots, x_{t-1}) \\ &+ \log p(x_t|x_{t+1}, x_{t+1}, \cdots, x_T))\end{aligned} \tag{4}$$

Different from bidirectional language model, BERT utilizes the masked language model (MLM), which allows bi-directional self-supervised pre-training more efficiently. It randomly selects 15% tokens of the input text, in which 80% of them are replaced with the special token "[MASK]", 10% of them are replaced with other words in the vocabulary. The objective is to maximize the log-likelihood of ground-truth words in the selected positions with masked text sequence:

$$\mathcal{L}_{mlm} = -\sum_{\hat{x} \in m(x)} \log p(\hat{x}|X_m) \tag{5}$$

where X_m is the masked text sequence, $m(x)$ is the set of masked words.

Training corpora Most methods use the corpora in the general domain such as BooksCorpus [39] and Wikipedia. To address the out-of-vocabulary words, they save subwords rather than words in the vocabulary. The Byte-Pair Encoding (BPE) [40] or WordPiece [41] methods are generally used to generate the vocabulary with subwords.

Fine-tuning Self-supervised pre-training on large scale of corpus allows language models to memorize common sense and linguistic knowledge in pre-trained parameters and contextual representations, which can be adapted to downstream tasks via fine-tuning with task-specific objective and datasets. For down stream tasks, task specific inputs are firstly fed into pre-trained language models to yield contextual representations. Different tasks usually be formulated into different problems such as classification, regression and generation problem. Therefore, for different tasks, it requires to choose contextual representations in different levels and different task-specific layers stacked on the top of language models. For example, for extractive summarization task, we first get the sentence representations from PLMs and add the extra classification layer to predict labels of sentences. All parameters of PLMs and task-specific parameters are refined with the task-specific loss with supervised data.

2.3 Biomedical Language Models

Inspired by the great success of PLMs on NLP tasks, much attention has been devoted to applying PLMs on tasks of biomedical domain including biomedical text summarization. Most advanced pre-trained language models, including BERT, variants of BERT, GPT3 [37], T5 [42] et al, are pre-trained on texts of general domain such as wikipedia and book corpus. Directly applying these language models to biomedical texts can be challenging. Since biomedical texts usually contain a lot of terminologies and compound words, which can not be covered by the vocabulary of PLMs pre-trained on texts of general domain.

To fill the gap, many pre-trained language models for biomedical domain such as BioBERT [16], and BlueBERT [43], ClinicalBERT [44] et al, have been proposed to further pre-train PLMs in general domain with biomedical texts. BioBERT [16] is the first biomedical language models that continues pre-training BERT on biomedical scientific texts including PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC). BlueBERT [43] further conduct continual pre-training on clinical texts: MIMIC-III [45], except from scientific texts. However, their vocabulary is still the same as that of BERT. Different from them, SciBERT [46] conducts pre-training and builds the novel vocabulary from scratch with scientific papers from mix-domain, in which 12% articles from the computer science domain and 82% articles from the biomedical domain. PubMedBERT [47] is pre-trained from scratch with scientific papers solely in biomedical domain. One can check the survey paper [20] for more details on pre-trained language models for biomedical domain.

3 Datasets

Unstructured biomedical texts used in text summarization methods involves in various types, including biomedical literature, electronic health records (EHRs), medical conversations, and medical questions as shown in Figure 3.

Biomedical Literature Biomedical scientific literature is an important source with rich clinical information. With the exponentially growing of scientific papers, developing automated summarization tools on biomedical articles has long attracted much attentions. Scientific papers are usually written by researchers and physicians. Compared with other texts such as social media texts, they have less noisy and are organized with standard sections, such as "Introduction", "Methods", "Results" et al.

PubMed [48] is one of the most commonly used dataset, for summarization of long biomedical texts. It consists of 133K scientific papers collected from the PubMed open access repositories[1]. To support development of multi-document summarization in biomedical domain, [15] build the RCT summarization dataset with 4,528 data samples searched from PubMed[2]. The input of each data sample includes titles and abstracts of papers describing randomized controlled trials (RCTs), while conclusion section of the systematic review from Cochrane[3] is treated as the target summary. Similarly, [14] developed the MSˆ2 for multi-document summarization of medical studies. It collected 470K papers from semantic scholar and 20K reviews that summarized these papers. [49] collected the CDSR dataset to support the task of lay language summarization of biomedical scientific reviews. It contains 7,805 abstracts pairs of biomedical scientific reviews, in which professional abstracts of systematic reviews as inputs, and their plain language abstracts as target summaries. SumPubMed [50] proposed recently, includes 33, 772 documents from Bio Med Central (BMC) of PubMed archive. [51], [52] proposed the subset from the large scientific corpus S2ORC [53], that includes 63,709 articles from the biological and biomedical domain. Moreover, to against COVID-19 pandemic, developing summarization systems on COVID-19 Open Research Dataset (CORD-19) [54] has attracted much attention. CORD-19 contains millions of papers related to COVID-19, SARS-CoV-2, and other coronaviruses. Details of these datasets are summarized in Table 1.

Electronic Health Records Electronic health records have been widely adopted by hospitals to store and manage medical information of patients, such as diagnostic codes, medications, laboratory results, clinical notes et al. They are written by professionals with professional language and specific structure. Different scientific papers that are generally free access, EHRs may have restrictions for publicly access due to the privacy issues. Several publicly available datasets have been released to support automated summarization of radiology reports. For radiology reports, their findings and background are considered as inputs, the aim is to generate impression section automatically that highlights key observations of reports. [55] collected the OpenI datasets containing 3,996 chest x-ray reports from hospitals within the Indiana Network. Compared with OpenI, MIMIC-CXR [56] is a larger publicly available dataset including 107,372 radiology reports from Beth Israel Deaconess Medical Center Emergency Department between 2011–2016. Statistics of these two datasets are shown in Table 2.

Medical Conversations Medical conversations between patients and doctors from online healthcare systems, has

1. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
2. https://pubmed.ncbi.nlm.nih.gov
3. https://www.cochranelibrary.com/

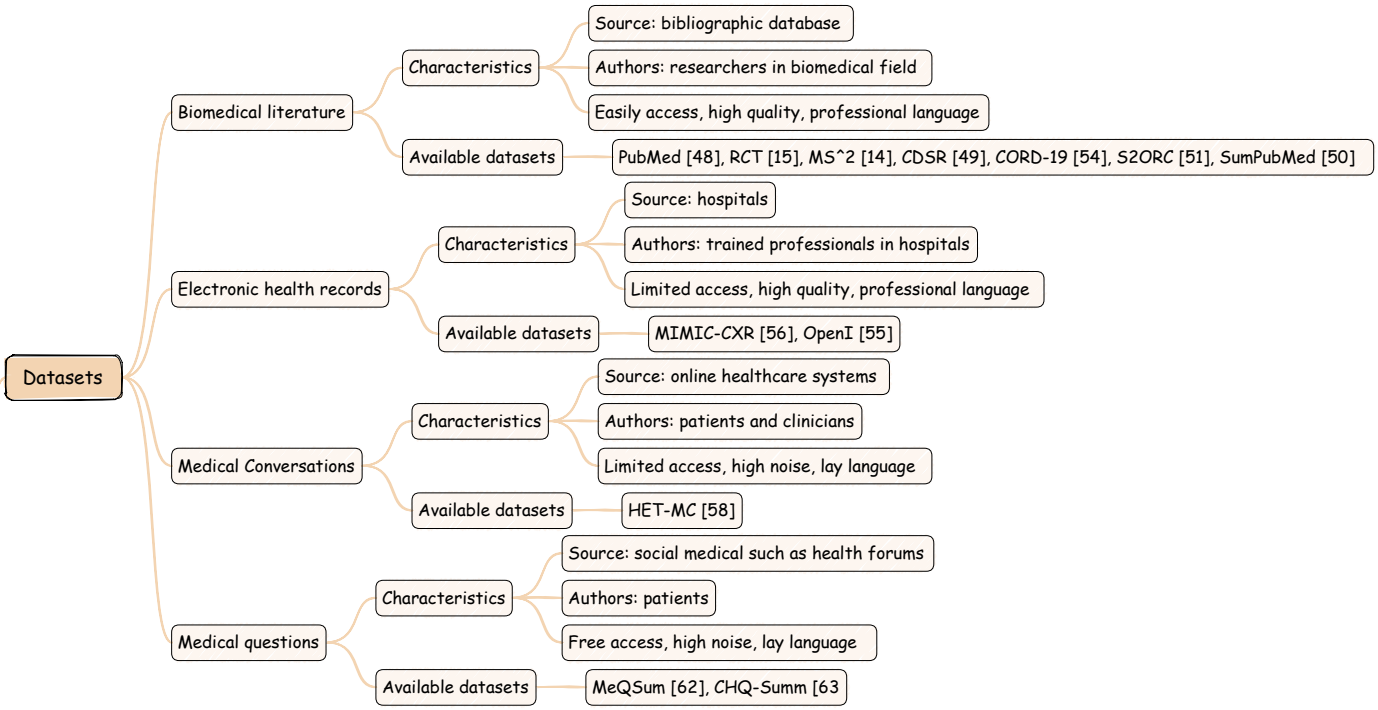


Fig. 3. Overview of datasets.

Dataset	Size	Content	Summarization Task	Access
PubMed [48]	133,215	Full contents of articles	Single	https://github.com/armancohan/long-summarization
RCT [15]	4,528	Titles and abstracts of articles	Multiple	https://github.com/bwallace/RCT-summarization-data
MS^2 [14]	470,402	Abstracts of articles	Multiple	https://github.com/allenai/ms2/
CDSR [49]	7,805	Abstracts of articles	Single	https://github.com/qiuweipku/Plain_language_summarization
SumPubMed [50]	33,772	Full contents of articles	Single	https://github.com/vgupta123/sumpubmed
S2ORC [51]	63,709	Full contents of articles	Single	https://github.com/jbshp/GenCompareSum
CORD-19 [54]	-	Full contents of articles	Single	https://github.com/allenai/cord19

TABLE 1
Biomedical literature datasets.

Dataset	Train	Val	Test
MIMIC-CXR [56]	122,014	957	1,606
OpenI [55]	2,400	292	576

TABLE 2
Statistics of EHR datasets.

Dataset	Train	Val	Test
MeQSum [62]	400	100	500
CHQ-Summ [63]	1,000	400	107

TABLE 3
Statistics of meidical question summarization datasets.

become an important source of medical information, with the increasing usage of telemedicine. Automated summarize key medical information on long medical conversations can save much time of doctors and improve healthcare efficiency. Medical conversations usually involve multi-turn interactions between two parties. The patients focus on asking questions and solution of their health problems and describing their symptoms. While doctors may ask detailed symptoms of patients and provide diagnostic suggestions. Similar to EHRs, accessing medical conversations at telemedicine platforms may have restrictions due to the privacy concerns. Moreover, it is time consuming and expensive to build the supervised data, since it requires professionals to write target summaries manually. Up to now, although several advanced methods with PLMs for medical coversation summarization have been proposed [57], [58], [59], [60], [61], publicly available datasets are limited. [58] proposed the Chinese medical coversation summarization dataset[4] with 109,850 conversations from the online health platfrom[5].

Medical Questions The consumer health questions produced by healthcare consumers in the web such as health forums, is another important data source of clinical information. To find trustworthy answers for their health questions, healthcare consumers can query the web with long natural language questions with peripheral details. The peripheral information is useless to find high-quality answers for health questions. Therefore, summarizing consumer health questions into concise text with salient information is quite

4. https://github.com/cuhksz-nlp/HET-MC
5. https://www.chunyuyisheng.com/

useful for improving efficiency of medical question answering. [62] build the MeQSum[6] corpus with 1,000 consumer health questions as inputs and their manually summaries from three medical experts. [63] introduced another dataset CHQ-Summ[7] most recently, that includes 1,507 consumer health questions and their summaries annotated by experts. Different from other texts such as biomedical papers with thousands words, consumer health question-summary pairs are short texts. Take CHQ-Summ for example, the average length of questions and their summaries are 200 words and 15 words respectively. Statistics of these two datasets are shown in Table 3.

4 PLMs for Biomedical Text Summarization

Many methods have been proposed to explore how to make better use of PLMs for the biomedical summarization task. According to types of output summary, numbers of input documents, and types of input documents, we categorize existing PLMs based methods into extractive summarization methods, abstractive summarization methods, multi-document summarization methods, medical dialogue summarization methods, and medical question summarization methods as shown in Figure 4. The ways of these methods to make use of PLMs can be summarized into feature based, fine-tuning based, and domain-adaption with fine-tuning based methods, as shown in Figure 5. The feature based methods utilize contextual representations from PLMs independently and directly without refining parameters of PLMs. The fine-tuning based methods usually take PLMs as the text encoder, and then fine-tune all parameters of PLMs along with task specific parameters. The fine-tuning with domain-adaption based methods, first conduct the domain-adaption for PLMs: continually pre-training PLMs with designed tasks on the target data, and then fine-tune the adapted PLMs along with task-specific layers. In next subsections, we will review and discuss these methods in more details.

4.1 Extractive Summarization Methods

For extractive summarization, existing methods can be categorized into supervised method and unsupervised method according to if they require labelled data. All supervised methods are fine-tuning based. [64] presents the ContinualBERT model for adaptive extractive summarization of covid-19 related literature. ContinualBERT trains two BERT models with continual learning in order to process texts online. [65] proposes BioBERTSum model for extractive summarization of biomedical literature. It uses token embedding, sentence embedding and position embedding to embed input texts, and then yields contextual representations of sentences with BioBERT. BioBERT and the extra classifier layer are fine-tuned with the cross-entropy loss. It proves the advantage of using domain-specific language model for biomedical texts, and outperforms the SOTA method BERTSum [66] that uses BERT in the general domain as encoder. [52] proposed the KeBioSum for the extractive summarization on biomedical literature. It proposes to incorporate the fine-grained medical knowledge into PLMs with the lightweight fine-tuning framework. It proves that although biomedical language models such as BioBERT, PubMedBERT can capture domain knowledge in some extent, fine-grained medical knowledge is still benefit to improve language models. [67] proposes the multi-head attention-based method for extractive summarization of clinical notes. It fine-tunes BERT on the task of predicting identify ICD-9 labels on the ICD-9 labeled MIMIC-III discharge notes. The attention scores of sentences from last layer of BERT model are used to select sentences.

Moreover, there are efforts that investigate unsupervised methods, which doesn't require labeled training data. They are usually the feature based that uses the contextual representations from PLMs as the extra features. [68] proposes the unsupervised extractive summarizer for biomedical texts based on hierarchical clustering and BERT. They conduct sentence clustering based on sentence representations from BERT and then select top-sentences from clusters with the ranking method. It shows better performance than traditional unsupervised methods such as TextLexAn[8]. [69] proposes the graph ranking based method for biomedical text summarization. They use the contextualized embeddings of BioBERT to represent sentences and build graphs for texts. The important sentences are identified with the graph ranking algorithm from text graphs. [51] presents unsupervised extractive summarization method GenCompareSum for biomedical literature. GenCompareSum uses the T5 generative model to generate key snippets for text sections, and selects important sentences with BERTScore [70] between key snippets and sentences. It outperforms traditional usupervised methods such as LexRank [71], TextRank [8] and also the SOTA supervised method BERTSum. [72] propose the radiology-specialized language model RadBERT that are pre-trained on millions of radiology reports. On the extractive summarization of radiology reports, it achieves better performance than other lanaguege models such as BERT, BioBERT.

4.2 Abstractive Summarization Methods

Most abstractive methods are based on fine-tuning pre-trained generative language models such as T5, GPT, BART [38]. [73] uses the GPT-2 [36] for abstractive summarization of COVID-19 medical research articles. They take keywords of articles as inputs and fine-tune GPT-2 on multi-tasks including the language modeling task and the multiple choice prediction task. [74] develops the encoder-decoder based framework for abstractive summarization of clinical notes. They propose to incorporate the contextual embeddings from BERT into Bi-LSTM based encoder. [49] uses the BART for automated lay language summarization of biomedical review articles. They conduct domain-adaption before fine-tuning, that pre-trains the BART model to reconstruct original PubMed abstracts with disrupted abstracts. Moreover, there are also methods based on the classical encoder-decoder framework, that take PLMs as text encoder and decoder. [75] proposes the radiology report summarizer that uses BioBERT as text encoder and randomly initialized transformer layers as decoder. [76] proposes SciBERT-based abstractive summarization model for COVID-19 scientific

6. https://github.com/abachaa/MeQSum
7. https://github.com/shwetanlp/Yahoo-CHQ-Summ
8. http://texlexan.sourceforge.net

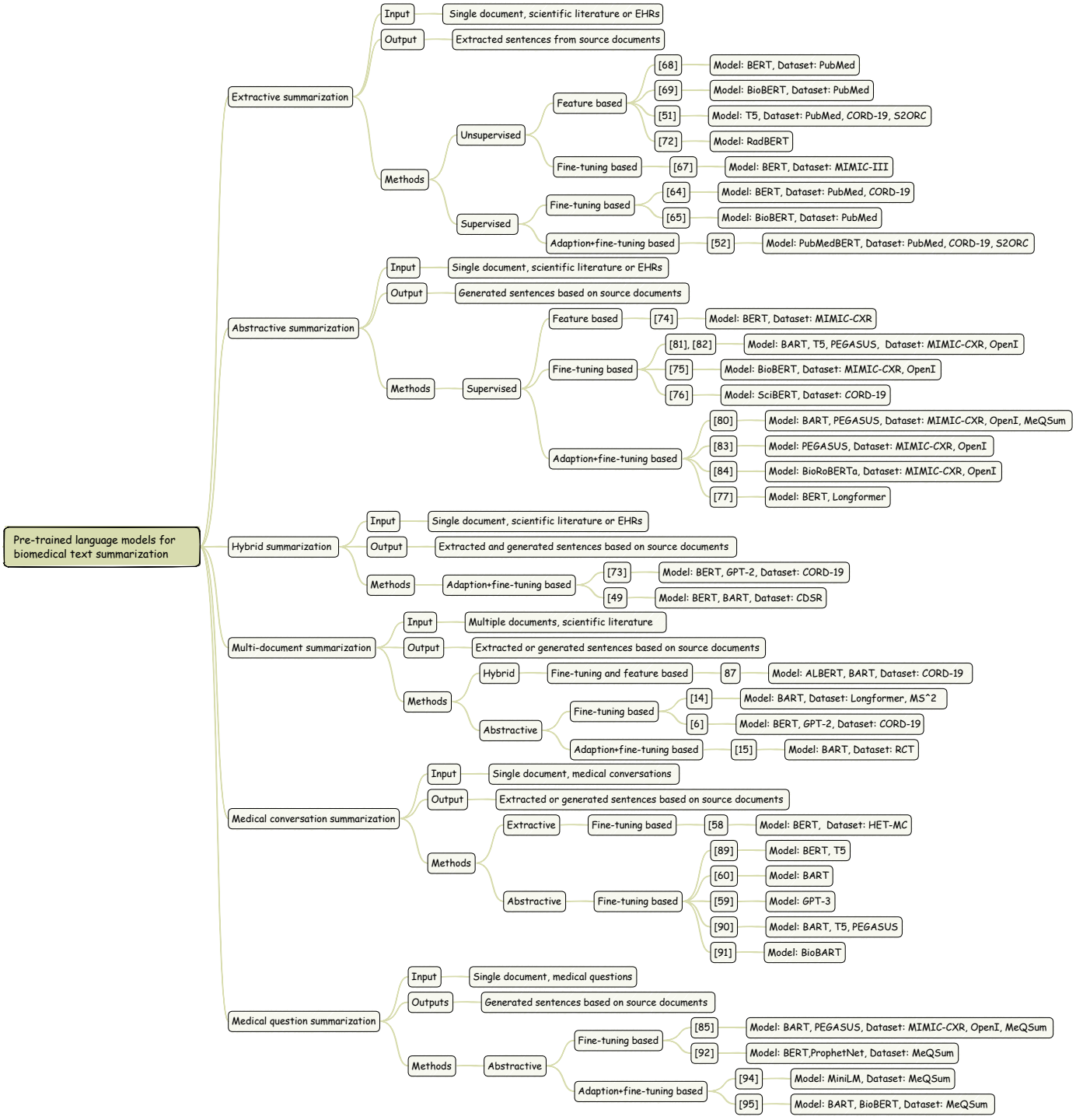
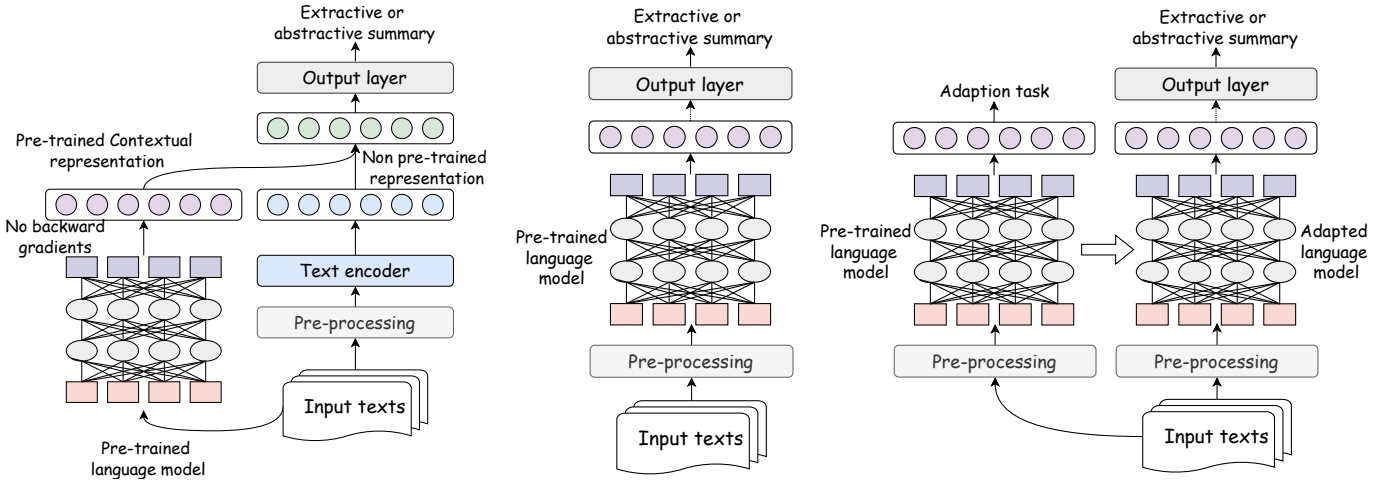


Fig. 4. Overview of methods.

papers, that uses the linguistic information of word co-occurrence encoded by graph attention network to enrich the SciBERT encoder. [77] presents the abstractive summarizer for patient hospitalisation histories, that uses Longformer [78] as the encoder and BERT as the decoder. They propose to pre-train BERT and Longformer with the masked language task on the hospitalisation history dataset before task specific fine-tuning.

Moreover, the MEDIQA 2021 Shared Task [79] at the BioNLP 2021 workshop includes the abstractive summarization task for radiology reports. Among 14 participated teams, 6 of them [80], [81], [82], [83], [84], [85] use the pre-trained language models such as BERT, BART, PEGASUS [86]. They find that the best performance is achieved by fin-tuning PEGASUS and BART. Moreover, they report that adapting PEGASUS on the PubMed corpus can lead to worse performance, which may due to the gap between biomedical literature and medical reports.



(a) The framework of feature based method. (b) The framework of fine-tuning based method. (c) The framework of adaption + fine-tuning based method.

Fig. 5. Comparison of different strategy on using pre-trained language models.

4.3 Multi-document Summarization Methods

Multi-document summarization that generates comprehensive summary for multiple documents is an valuable task for real applications of the biomedical domain, such as systematic reviews. Most methods are based on the encoder-decoder framework, and concatenate multiple documents to formulate the single final input. [15] proposes the multi-document abstractive summarization models based on BART for randomized controlled trials (RCTs). They adapt the BART with the domain-specific pre-training strategy of generating summaries from full-text articles before fine-tuning. They also use the "decoration" strategy to explicitly inform key trial attributes (the "PICO" elements) of input articles. [14] develops the BART based method for multi-document summarization of medical studies. To encode multi-documents, they investigates two encoders. One is using multiple BART encoders to encode multi-documents separately. Another one is using LongformerEncoderDecoder (LED) [78], which can encode long inputs up to 16K tokens.

Moreover, there are multi-document summarization methods to help information retrieval of COVID-19 literature. [6] designs the parallel encoder-decoder framework for abstractive summarization of multiple COVID-19 articles, that uses BERT as encoder and GPT-2 as decoder. [87] proposes the query-focused multi-document summarizer for COVID-19 articles, that is able to generate abstractive and extractive summary based on user queries. They fintune BART for multi-document abstractive summarization. For extractive summarization, they use ALBERT [88] to generate sentence representations and calculate the cosine similarity between sentences and queries to select important sentences.

4.4 Medical Dialogue Summarization Methods

Fine-tuning PLMs for summarization of medical conversation faces several challenges including: limited labeled data and long transcripts. Several efforts are proposed to deal with these issues. [58] proposes the hierarchical encoder-tagger (HET) model for extractive summarization of medical conversation, that includes token-level and utterance-level encoders to encode input long transcripts. They use BERT as the token-level encoder. [89] develops CLUSTER2SENT, a extractive-abstractive hybrid method on doctor-patient conversations to generate SOAP notes (long semi-structured clinical summaries). T5 model is used in the abstractive module of CLUSTER2SENT. [60] leverages BART model for automatic summarization of doctor-patient conversations. They propose the multistage fine-tuning strategy to address the input length limitation of BART. They also find that fine-tuning BART can generate summaries in good quality even with limited training data.

To overcome the problem of limited training data, methods are proposed to use the few-shot learner such as GPT-3 or few-shot fine-tuning strategy. [59] integrates medical knowledge and GPT-3 for medical dialogue summarization. They consider GPT-3 as the summary generator and choose the best summary that captures the most medical concepts. [90] explores fine-tuning PLMs including BART, T5, PEGASUS [86] with zero-shot and few-shot learning strategy for medical dialogue summarization with small training data. They find that BART achieves the best performance among these PLMs. Moreover, [91] develops the first generative pre-trained language model BioBART in the biomedical domain, which has shown better performance on medical dialogue summarization than BART.

4.5 Medical Question Summarization Methods

[62] is the earliest work that proposes the task of summarization of consumer health questions. To facilitate the development of the task, the question summarization task is included at the MEDIQA 2021 Shared Task [79]. All methods submitted to the question summarization task are based on the pre-trained language models. The best performance achieved by the ensemble model [85] that re-ranks summary outputs of multiple advanced generative language models

Paper	Category	Strategy	Model	Training	Data
ContinualBERT [64]	extractive	fine-tuning	BERT	supervised	PubMed, CORD-19
BioBERTSum [65]	extractive	fine-tuning	BioBERT	supervised	PubMed
KeBioSum [52]	extractive	adaption+fine-tuning	PubMedBERT	supervised	PubMed, CORD-19, S2ORC
[67]	extractive	fine-tuning	BERT	unsupervised	MIMIC-III
[68]	extractive	feature-base	BERT	unsupervised	PubMed
[69]	extractive	feature-base	BioBERT	unsupervised	PubMed
GenCompareSum [51]	extractive	feature-base	T5	unsupervised	PubMed, CORD-19, S2ORC
RadBERT [72]	extractive	feature-base	RadBERT	unsupervised	-
[73]	hybrid	adaption+fine-tuning	BERT,GPT-2	supervised	CORD-19
[74]	abstractive	feature-base	BERT	supervised	MIMIC-CXR
[49]	hybrid	adaption+fine-tuning	BERT, BART	supervised	CDSR
[80]	abstractive,question	adaption+fine-tuning	BART,PEGASUS	supervised	MIMIC-CXR,OpenI,MeQSum
[81]	abstractive	fine-tuning	BART,T5,PEGASUS	supervised	MIMIC-CXR,OpenI
[82]	abstractive	fine-tuning	BART,T5,PEGASUS	supervised	MIMIC-CXR,OpenI
[83]	abstractive	adaption+fine-tuning	PEGASUS	supervised	MIMIC-CXR,OpenI
[84]	abstractive	adaption+fine-tuning	BioRoBERTa [96]	supervised	MIMIC-CXR,OpenI
[75]	abstractive	fine-tuning	BioBERT	supervised	MIMIC-CXR,OpenI
[76]	abstractive	fine-tuning	SciBERT	supervised	CORD-19
[77]	abstractive	adaption+fine-tuning	BERT,Longformer	supervised	-
[15]	abstractive,multi-doc	adaption+fine-tuning	BART	supervised	RCT
[14]	abstractive,multi-doc	fine-tuning	BART,Longformer	supervised	MSˆ2
[6]	abstractive,multi-doc	fine-tuning	BERT,GPT-2	supervised	CORD-19
[87]	hybrid,multi-doc	fine-tuning,feature-base	ALBERT,BART	un+supervised	CORD-19
HET [58]	extractive,dialogue	fine-tuning	BERT	supervised	HET-MC
CLUSTER2SENT [89]	abstractive,dialogue	fine-tuning	BERT,T5	supervised	-
[60]	abstractive,dialogue	fine-tuning	BART	supervised	-
[59]	abstractive,dialogue	fine-tuning	GPT-3	supervised	-
[90]	abstractive,dialogue	fine-tuning	BART,T5, PEGASUS	supervised	-
BioBART [91]	abstractive,dialogue	fine-tuning	BioBART	supervised	-
[85]	abstractive,question	fine-tuning	BART,T5,PEGASUS	supervised	MeQSum,MIMIC-CXR,OpenI
[92]	abstractive,question	fine-tuning	BERT,ProphetNet	supervised	MeQSum
[94]	abstractive,question	adaption+fine-tuning	Minilm [97]	supervised	MeQSum
[95]	abstractive,question	adaption+fine-tuning	BART,BioBERT	supervised	MeQSum

TABLE 4
Overview of Methods. "-" means datasets that are not released.

including BART, T5 and PEGASUS. [92] presents the reinforcement learning based framework for abstractive summarization of medical questions. They proposes two reward function: the Question-type Identification Reward (QTR) and Question-focus Recognition Reward (QFR), which are optimized via learning optimal policy defined by BERT. They show that generative language model ProphetNet [93] with the proposed reward functions has better performance than other PLMs including T5, BART, and PEGASUS. [94] investigates to incorporate the knowledge of "question-focus" and "question-type" with PLMs for abstractive summarization of consumer health question. To induce PLMs to capture these knowledge, they adapt PLMs with designed Cloze tasks. [95] presents the multi-task learning and data augmentation method on medical question summarization and recognizing question entailment (RQE) for medical question understanding. They prove that the multi-task learning between question summarization and RQE is able to increase performance of PLMs including BART and BioBERT.

4.6 Discussion

We compare all methods in Table 4, and make a further discussion in this section.

Extractive or abstractive Most methods for multi-document summarization, medical dialogue summarization and medical question summarization are abstractive summarization methods. Most extractive methods are used for biomedical literature. Three methods [49], [73], [87] are hybrid approaches that consider both abstractive and extractive summarization.

Supervised or unsupervised All abstractive summarization methods are supervised approaches based on fine-tuning. Extractive methods can be categorized into supervised and unsupervised methods. All supervised extractive methods are fine-tuning based, while unsupervised extractive methods are feature based.

Choice of language models For extractive summarization, most methods use the classical language model: BERT. A part of methods [52], [65], [72] leverage language models for biomedical domain including BioBERT, PubMedBERT and RadBERT. For abstractive summarization, generative language models: BART, T5, PEGASUS are used by most methods. The Longformer are used by several methods [14], [77] to encode long inputs.

Fine-tuning or feature-base All abstractive summarization methods are fine-tuning based methods. Several extractive methods [51], [68], [69], [72], [87] are feature based and are all unsupervised methods. To fill the gap between PLMs in general domain and biomedical texts, several methods [14], [15], [49], [52], [73], [94] propose to adapt language models with masked language task or other designed tasks to capture domain-specific knowledge before fine-tuning.

5 Evaluations

Evaluating summaries of biomedical texts is more challenging than general document summarization, since biomedical texts are much different from general texts in length and

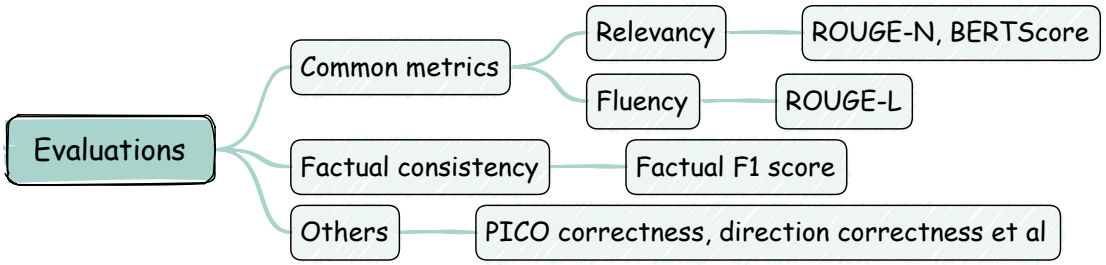


Fig. 6. Overview of evaluations.

structure. They are more technical and complex. Existing methods focus on evaluating several aspects including: 1) relevancy: how relevance of generated summaries with gold summaries, 2) fluency: how fluent generated summaries are, 3) factuality: how factual agreement of generated summaries with original documents. Figure 6 shows the overview of evaluations.

Common metrics Manually evaluating performance of summarization methods is time-consuming and expensive. Automatic evaluation metrics have been proposed and widely used to evaluate summarization systems efficiently. Most of them are based on measuring similarity between generated summaries of automatic approaches and gold summaries. Similar to the general domain, ROUGE [98] is the most widely used metric for biomedical summarizers. The most commonly used metrics in ROUGE family include:

- ROUGE-N: N-gram overlap between generated summaries of summarizers and gold summaries, which are used to evaluate the relevance between them.
- ROUGE-L: the longest common subsequences between generated summaries of summarizers and gold summaries, which are used to evaluate fluency of generated summaries.

Most methods report F1 scores of ROUGE-1, ROUGE-2, and ROUGE-L. However, ROUGE metrics are not effective enough due to only rely on the shallow lexical overlaps. They don't consider the paraphrasing and terminology variations when measuring similarity. To address it, one important work is [70], that proposes the novel metric BERTScore, that calculates similarity between two sentences with the sum of cosine similarities between contextual embeddings of tokens from pre-trained language models.

Factual consistency Factual correctness of generated summaries is a critical measurement for the real application of automatic systems, especially in the biomedical domain. Compared with extractive methods, it is reported [99] that abstractive methods are struggle to generate factual correct summaries. [100] proposes the factual F1 score to evaluate the factual correctness of generated summaries of radiology reports. They propose to use the CheXbert labeler [101] to yield the binary presence values of disease variables of generated summaries and references, and then calculate the overlap of yielded binary presence values between them. [14] proposes the ΔEI metric to calculate the factual agreement of generated summaries and input medical studies. They propose to calculate the Jensen-Shannon Distance (JSD) between distributions of generated summaries and input medical studies on three directions (increase, decrease, no change) of reported directionality.

Other studies [102] studies the correlation between humman evaluation and 18 automatic evaluation metrics including text overlap metrics: ROUGE, CHRF [103], embedding metrics BertScore, factual F1 score et al, on generated clinical consultation notes. They find that simple character-based metrics such as character-based Levenshtein distance can be more effective than other complex metrics such as BERTScore, and the choice of human references can largely influence the performance of automatic evaluation metrics. [104] proposes a objective human evaluation based on counting medical facts for generated summaries of medical reports. [105] proposes a new human evaluation approach for generated summaries of multiple biomedical documents, concentrating on quality dimensions including factuality, PICO correctness, direction correctness et al. They show that existing automatic metrics such as ROUGE are not effective on evaluating factual consistency of generated summaries.

6 Limitations and Future Directions

In this section, we summarize limitations of existing methods and promising future directions as shown in Figure 7.

6.1 Limitations

Although PLMs based methods have greatly boosted the performance of biomedical summarization, there are still several limitations.

Encoding long biomedical texts PLMs have the limitation on the token length of input documents [78] due to the high time and memory consumption of attention computations in Transformer. Most PLMs based summarization methods directly truncate input documents and only take their first 512 tokens following methods in general domain. However, biomedical texts such as biomedical scientific papers, usually have thousands of tokens. The truncating operation losses useful information in truncated contents of input documents. Moreover, it also leads to the losses of long-range dependencies on long biomedical documents. Therefore, it is still a limitation for existing methods to encode long biomedical texts efficiently.

Incorporating domain-specific knowledge Domain-specific knowledge is critical for understanding biomedical texts. Vocabularies, taxonomies and ontologies such as UMLS [106] are important sources of biomedical knowledge. While existing methods with PLMs is able to capture lexical knowledge in biomedical texts, they have no

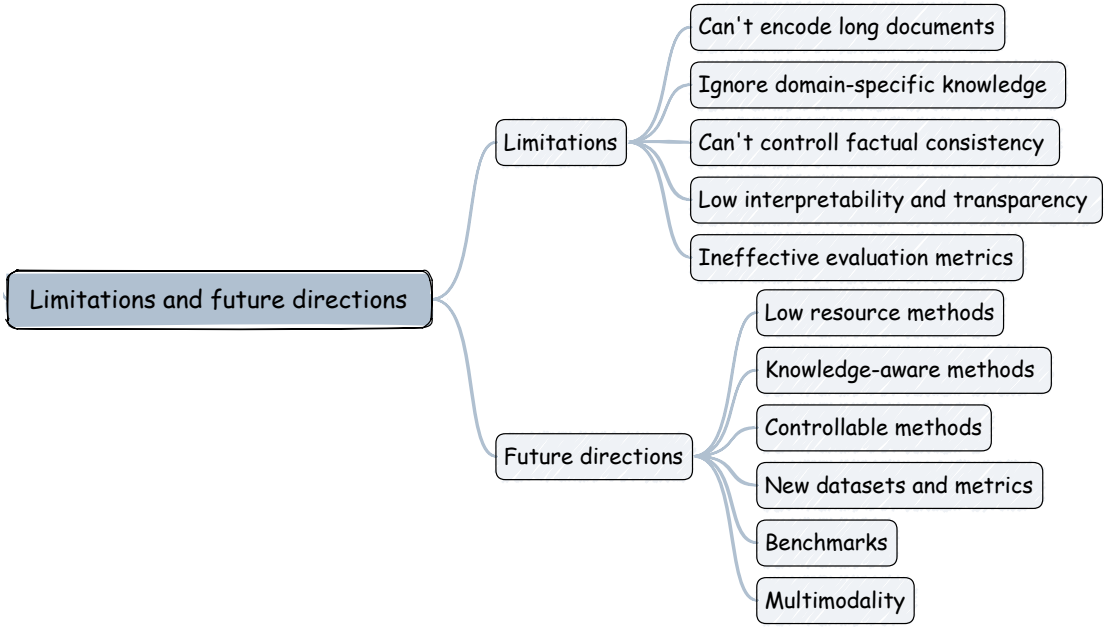


Fig. 7. Overview of limitations and future Directions.

knowledge of words or entities that have particular domain-specific importance, and their relations. It is still a challenge for existing methods to capture knowledge of sources such as biomedical concepts, relations between concepts, and lexicographic information et al.

Controlling factual consistency Factual correctness of generated summaries is especially important for real application of automatic biomedical summarizers. However, existing abstractive methods are encouraged to reconstruct gold summaries freely without word constraints on text generation. They are tend to generate summaries that fabricates facts of original inputs due to freely rephrasing [100], [107], which may cause medical errors. It is a challenge for existing methods to control factual consistency when fine-tuning PLMs.

Interpretability and transparency Similar to other deep learning methods, PLMs based methods have the well-know interpretability problem due to black-box nature of PLMs. For users such as clinicians, they are hard to explain how and why models select specific words or sentences to yield the final summaries. If errors are consistently made by the model, it is hard for users to know why things go wrong. The explainability and transparency of models such as inner mechanisms of their algorithms, are important in constructing reliable applications for users [108].

Evaluations Objective and comprehensive evaluation metrics are important to evaluate summarization methods efficiently and reliably. Most existing methods only use the ROUGE metrics to evaluate their models automatically similar to methods in the general domain. However, it has reported that ROUGE is far from reflecting quality of generated summaries accurately such as factual correctness, key finding directions. Although there are efforts that explores objective humman evaluation metrics for medical studies [14], [105], it is still lack of accurate automatic evaluation metric which is compatible with human.

Overall, performances of existing methods are still far from desirable. We believe more efforts should be proposed to address these limitations.

6.2 Future Directions

In this section, we further discuss promising future directions, which we hope can provide guidelines for future research.

Low resources In biomedical domain, it usually requires professionals such as clinicians to conduct data annotation, which is expensive and time-consuming. Although PLMs based methods are able to transfer semantic information of large scale of unlabeled data via fine-tuning PLMs, most of them are still supervised methods and require large amount of labelled data. We believe more efforts should be proposed in the future to explore unsupervised, few-shot learning, and data augmentation techniques for low resource biomedical summarization.

Incorporating extra knowledge PLMs for general domain and biomedical domain, are shown to be able to capture common sense knowledge and biomedical knowledge in certain extent. Although they can generate summaries that are fluent or grammatically correct, it proves that most of their generated summaries are illogical or have factual errors [14], [105]. Therefore, limited knowledge captured by PLMs is hard to support the model to generate desirable summaries. It is expect that more knowledge-aware models can be proposed to incorporate extra domain-specific knowledge such as knowledge base UMLS, to improve summarization generation.

Controllable generation Existing methods generally yield summaries that ignore users' preferences. We believe more efforts should be developed for controlled summarization of biomedical texts, that meet expectations and requirements of users. Methods are expected to control

several attributes of generated summaries, such as length, readability, text style et al.

New datasets and metrics Although there are several biomedical literature datasets, there are limited publicly available medical conversation and question datasets due to annotation and privacy issues. Moreover, there are rare attentions on developing auto matic evaluation metrics that are suitable for biomedical texts. It is expect that more datasets and evaluation metrics can be proposed to support development and evaluation of the task. Moreover, it is also promising to use techniques such as federated learning [109] to allow developing of models while keep training data in the private side.

Benchmarks To facilitate develpment of biomedical NLP, attempts have been made to create NLP benchmark in the biomedical domain such as BLUE [43], GLUE [47], that include relation extraction task, text classification task et al. However, none of existing benchmark includes the biomedical text summarization task. Considering variety types of biomedical texts including scientific papers, EHRs, conversations, questions, and categories of tasks including extractive, abstractive and multi-documents summarization, we believe it is necessary to build the unified benchmark to support development and fair evaluations of proposed methods.

Multimodality In biomedical domain, there are rich multimodal medical dataset such as radiology reports, and associated x-rays. However, most existing methods only take biomedical texts themselves as inputs. It reports that visual features can improve performance of text generation [110]. It expects that multimodal summarization methods can draw much attention in the future.

7 Conclusion

In this survey, we make a comprehensive overview of biomedical text summarization with large scale pre-trained language models . We systematically review recent approaches that are based on PLMs, benchmark datasets, evaluations of the task. We categorize and compare recent approaches according to types of input documents, numbers of input documents, and types of output summaries. Finally, we highlight limitations of existing methods and suggest potential directions for future research. We hope the paper can be a timely survey to help future researchers.

Acknowledgments

This research is supported by the Alan Turing Institute and the Biotechnology and Biological Sciences Research Council (BBSRC), BB/P025684/1.

References

[1] Z. Lu, "Pubmed and beyond: a survey of web tools for searching biomedical literature," *Database*, vol. 2011, 2011.

[2] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[3] M. Maybury, *Advances in automatic text summarization*. MIT press, 1999.

[4] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, and G. Del Fiol, "Text summarization in the biomedical domain: a systematic review of recent research," *Journal of biomedical informatics*, vol. 52, pp. 457–467, 2014.

[5] P. Przybyła, A. J. Brockmeier, G. Kontonatsios, M.-A. Le Pogam, J. McNaught, E. von Elm, K. Nolan, and S. Ananiadou, "Prioritising references for systematic reviews with robotanalyst: a user study," *Research synthesis methods*, vol. 9, no. 3, pp. 470–488, 2018.

[6] A. Esteva, A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev, and R. Socher, "Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–9, 2021.

[7] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *Journal of biomedical informatics*, vol. 42, no. 5, pp. 760–772, 2009.

[8] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[9] J. M. Conroy and D. P. O'leary, "Text summarization via hidden markov models," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 406–407.

[10] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 484–494.

[11] C. Friedman, P. Kra, and A. Rzhetsky, "Two biomedical sublanguages: a description based on the theories of zellig harris," *Journal of biomedical informatics*, vol. 35, no. 4, pp. 222–235, 2002.

[12] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[13] A. Roberts, C. Raffel, and N. Shazeer, "How much knowledge can you pack into the parameters of a language model?" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5418–5426.

[14] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, "Ms^2: Multi-document summarization of medical studies," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7494–7513.

[15] B. C. Wallace, S. Saha, F. Soboczenski, and I. J. Marshall, "Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization," *AMIA Summits on Translational Science Proceedings*, vol. 2021, p. 605, 2021.

[16] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[17] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos, "Summarization from medical documents: a survey," *Artificial intelligence in medicine*, vol. 33, no. 2, pp. 157–177, 2005.

[18] R. Pivovarov and N. Elhadad, "Automated methods for the summarization of electronic health records," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 938–947, 2015.

[19] M. Wang, M. Wang, F. Yu, Y. Yang, J. Walker, and J. Mostafa, "A systematic review of automatic text summarization for biomedical literature and ehrs," *Journal of the American Medical Informatics Association*, vol. 28, no. 10, pp. 2287–2297, 2021.

[20] B. Wang, Q. Xie, J. Pei, P. Tiwari, Z. Li *et al.*, "Pre-trained language models in biomedical domain: A systematic survey," *arXiv preprint arXiv:2110.05006*, 2021.

[21] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[22] H. Lin and V. Ng, "Abstractive summarization: A survey of the state of the art," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9815–9822.

[23] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 1995, pp. 68–73.

[24] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015*

Conference on Empirical Methods in Natural Language Processing, 2015, pp. 379–389.

[25] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[26] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[28] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[29] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Jun. 2018, pp. 2227–2237.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[32] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1756–1765.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[35] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training."

[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners."

[37] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[38] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[39] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[40] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), 2016, pp. 1715–1725.

[41] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.

[42] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[43] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 58–65.

[44] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, W. Redmond, and M. B. McDermott, "Publicly available clinical bert embeddings," *NAACL HLT 2019*, p. 72, 2019.

[45] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[46] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3615–3620.

[47] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[48] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," in *Proceedings of NAACL-HLT*, 2018, pp. 615–621.

[49] Y. Guo, W. Qiu, Y. Wang, and T. Cohen, "Automated lay language summarization of biomedical scientific reviews," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 160–168.

[50] V. Gupta, P. Bharti, P. Nokhiz, and H. Karnick, "Sumpubmed: Summarization dataset of pubmed scientific articles," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, 2021, pp. 292–303.

[51] J. Bishop, Q. Xie, and S. Ananiadou, "Gencomparesum: a hybrid unsupervised summarization method using salience," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 220–240.

[52] Q. Xie, J. A. Bishop, P. Tiwari, and S. Ananiadou, "Pre-trained language models with domain knowledge for biomedical extractive summarization," *Knowledge-Based Systems*, p. 109460, 2022.

[53] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. S. Weld, "S2orc: The semantic scholar open research corpus," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4969–4983.

[54] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney *et al.*, "Cord-19: The covid-19 open research dataset," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

[55] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[56] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, pp. 1–8, 2019.

[57] S. Enarvi, M. Amoia, M. D.-A. Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto *et al.*, "Generating medical reports from patient-doctor conversations using sequence-to-sequence models," in *Proceedings of the first workshop on natural language processing for medical conversations*, 2020, pp. 22–30.

[58] Y. Song, Y. Tian, N. Wang, and F. Xia, "Summarizing medical conversations via identifying important utterances," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 717–729.

[59] B. Chintagunta, N. Katariya, X. Amatriain, and A. Kannan, "Medically aware gpt-3 as a data generator for medical dialogue summarization," in *Machine Learning for Healthcare Conference*. PMLR, 2021, pp. 354–372.

[60] L. Zhang, R. Negrinho, A. Ghosh, V. Jagannathan, H. R. Hassanzadeh, T. Schaaf, and M. R. Gormley, "Leveraging pretrained models for automatic summarization of doctor-patient conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3693–3712.

[61] W.-w. Yim and M. Yetisgen-Yildiz, "Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization," in *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, 2021, pp. 10–20.

[62] A. B. Abacha and D. Demner-Fushman, "On the summarization of consumer health questions," in *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics, 2019, pp. 2228–2234.

[63] S. Yadav, D. Gupta, and D. Demner-Fushman, "Chq-summ: A dataset for consumer healthcare question summarization," *arXiv preprint arXiv:2206.06581*, 2022.

[64] J. W. Park, "Continual bert: Continual learning for adaptive extractive summarization of covid-19 literature," *arXiv preprint arXiv:2007.03405*, 2020.

[65] Y. Du, Q. Li, L. Wang, and Y. He, "Biomedical-domain pre-trained language model for extractive summarization," *Knowledge-Based Systems*, vol. 199, p. 105964, 2020.

[66] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3730–3740.

[67] N. Kanwal and G. Rizzo, "Attention-based clinical note summarization," in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 2022, pp. 813–820.

[68] M. Moradi, G. Dorffner, and M. Samwald, "Deep contextualized embeddings for quantifying the informative content in biomedical text summarization," *Computer methods and programs in biomedicine*, vol. 184, p. 105117, 2020.

[69] M. Moradi, M. Dashti, and M. Samwald, "Summarization of biomedical articles using domain-specific word embeddings and graph ranking," *Journal of Biomedical Informatics*, vol. 107, p. 103452, 2020.

[70] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2019.

[71] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[72] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu, "Radbert: Adapting transformer-based language models to radiology," *Radiology: Artificial Intelligence*, p. e210258, 2022.

[73] V. Kieuvongngam, B. Tan, and Y. Niu, "Automatic text summarization of covid-19 medical research articles using bert and gpt-2," *arXiv preprint arXiv:2006.01997*, 2020.

[74] S. S. Gharebagh, N. Goharian, and R. Filice, "Attend to medical ontologies: Content selection for clinical abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1899–1905.

[75] H. Jingpeng, L. Zhuo, C. Zhihong, L. Zhen, W. Xiang, and C. Tsung-Hui, "Graph enhanced contrastive learning for radiology findings summarization," in *Proceedings of Association for Computational Linguistics (ACL)*, vol. 2, 2022.

[76] X. Cai, S. Liu, L. Yang, Y. Lu, J. Zhao, D. Shen, and T. Liu, "Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers," *Journal of Biomedical Informatics*, vol. 127, p. 103999, 2022.

[77] A. Yalunin, D. Umerenkov, and V. Kokh, "Abstractive summarization of hospitalisation histories with transformer networks," *arXiv preprint arXiv:2204.02208*, 2022.

[78] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[79] A. B. Abacha, Y. M'rabet, Y. Zhang, C. Shivade, C. Langlotz, and D. Demner-Fushman, "Overview of the mediqa 2021 shared task on summarization in the medical domain," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 74–85.

[80] L. Xu, Y. Zhang, L. Hong, Y. Cai, and S. Sung, "Chichealth@ mediqa 2021: Exploring the limits of pre-trained seq2seq models for medical summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 263–267.

[81] W. Zhu, Y. He, L. Chai, Y. Fan, Y. Ni, G. Xie, and X. Wang, "paht_nlp@ mediqa 2021: Multi-grained query focused multi-answer summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 96–102.

[82] R. Kondadadi, S. Manchanda, J. Ngo, and R. McCormack, "Optum at mediqa 2021: Abstractive summarization of radiology reports using simple bart finetuning," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 280–284.

[83] S. Dai, Q. Wang, Y. Lyu, and Y. Zhu, "Bdkg at mediqa 2021: System report for the radiology report summarization task," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 103–111.

[84] D. Mahajan, C.-H. Tsou, and J. J. Liang, "Ibmresearch at mediqa 2021: Toward improving factual correctness of radiology report abstractive summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 302–310.

[85] Y. He, M. Chen, and S. Huang, "damo_nlp at mediqa 2021: knowledge-based preprocessing and coverage-oriented reranking for medical question summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 112–118.

[86] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[87] D. Su, Y. Xu, T. Yu, F. B. Siddique, E. Barezi, and P. Fung, "Caire-covid: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

[88] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2019.

[89] K. Krishna, S. Khosla, J. P. Bigham, and Z. C. Lipton, "Generating soap notes from doctor-patient conversations using modular summarization techniques," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4958–4972.

[90] D. F. Navarro, M. Dras, and S. Berkovsky, "Few-shot fine-tuning sota summarization models for medical dialogues," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 2022, pp. 254–266.

[91] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "Biobart: Pretraining and evaluation of a biomedical generative language model," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 97–109.

[92] S. Yadav, D. Gupta, A. B. Abacha, and D. Demner-Fushman, "Reinforcement learning for abstractive question summarization with question-aware semantic rewards," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 249–255.

[93] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "Prophetnet: Predicting future n-gram for sequence-to-sequencepre-training," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2401–2410.

[94] S. Yadav, D. Gupta, A. B. Abacha, and D. Demner-Fushman, "Question-aware transformer models for consumer health question summarization," *Journal of Biomedical Informatics*, vol. 128, p. 104040, 2022.

[95] K. Mrini, F. Dernoncourt, S. Yoon, T. Bui, W. Chang, E. Farcas, and N. Nakashole, "A gradually soft multi-task and data-augmented approach to medical question understanding," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 1505–1515.

[96] S. Rothe, S. Narayan, and A. Severyn, "Leveraging pre-trained checkpoints for sequence generation tasks," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 264–280, 2020.

[97] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776–5788, 2020.

[98] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[99] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz, "Learning to summarize radiology findings," in *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, 2018, pp. 204–213.

[100] Y. Zhang, D. Merck, E. B. Tsai, C. D. Manning, and C. Langlotz, "Optimizing the factual correctness of a summary: A study of summarizing radiology reports," in *ACL*, 2020.

[101] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. Lungren, "Combining automatic labelers and expert annotations for accurate radiology report labeling using bert," in *Proceedings*

of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1500–1519.

[102] F. Moramarco, A. P. Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, and A. Savkov, "Human evaluation and correlation with automatic metrics in consultation note generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5739–5754.

[103] M. Popović, "chrf: character n-gram f-score for automatic mt evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015, pp. 392–395.

[104] F. Moramarco, D. Juric, A. Savkov, and E. Reiter, "Towards objectively evaluating the quality of generated medical summaries," in *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, 2021, pp. 56–61.

[105] J. Otmakhova, K. Verspoor, T. Baldwin, and J. H. Lau, "The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5098–5111.

[106] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[107] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9332–9346.

[108] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[109] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.

[110] J.-B. Delbrouck, C. Zhang, and D. Rubin, "Qiai at mediqa 2021: Multimodal radiology report summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 285–290.